International Research on Big Data and Computer Technology
Vol 10 No 1 – April 2026
ISSN 2797-4278
DOI: 10.53514/ir.v10i1.725

Penerapan *K-Means Clustering* Untuk Pengelompokan Modalitas Pencitraan Medis Pada Dataset *Imaging Data Commons*

[1]Dea Anggraini, [1]Angelina S. Saragih, [1]Bicanro G Panjaitan

[1]*Universitas Negeri Medan,*
anggrainidea2006@gmail.com, angelina25saragih@gmail.com, bicanropanjaitan@gmail.com

ABSTRAK

Distribusi modalitas pencitraan medis dalam dataset Imaging Data Commons (IDC) yang berskala petabyte belum pernah dianalisis secara sistematis menggunakan pendekatan data mining, padahal ketimpangan distribusinya berpotensi memengaruhi kualitas pengembangan model kecerdasan buatan di bidang onkologi. Penelitian ini bertujuan mengelompokkan modalitas pencitraan medis berdasarkan karakteristik distribusi data menggunakan algoritma K-Means Clustering yang diakses melalui Google BigQuery. Data diproses melalui tahapan preprocessing meliputi pembersihan, penyaringan noise, dan normalisasi MinMaxScaler. Nilai K optimal ditentukan melalui Elbow Method menghasilkan K=3. Hasil clustering membentuk tiga kelompok: Cluster 2 (CT, 62,22%), Cluster 1 (MR, 32,41%), dan Cluster 0 (20 modalitas minor, 5,37%), dengan Silhouette Score 0,7823 yang termasuk kategori cluster kuat. Penelitian ini mengungkap bahwa hubungan antara cakupan body part dan volume data bersifat eksponensial, serta dominasi CT dan MRI berpotensi menciptakan blind spot pada model AI medis apabila ketimpangan distribusi tidak ditangani sebelum pelatihan model.

Keyword: K-Means Clustering, Imaging Data Commons, Data Mining, Unsupervised Learning

1. PENDAHULUAN

Data mining merupakan proses komputasional untuk mengekstrak pola, informasi, dan pengetahuan tersembunyi dari kumpulan data berukuran besar secara otomatis. Menurut Aulia et al. (2024) data mining adalah proses analisis dan ekstraksi pola yang bermanfaat dari dataset dalam jumlah besar dengan pendekatan statistik, matematika, dan algoritma berbasis komputer yang dimanfaatkan untuk mengidentifikasi hubungan tersembunyi dan mendukung pengambilan keputusan di berbagai bidang seperti kesehatan, bisnis, dan teknologi. Ramadhani et al. (2025) menyatakan bahwa data mining telah berkembang menjadi salah satu cabang penting dalam bidang kecerdasan buatan, melibatkan teknik seperti machine learning, statistik, dan matematika untuk mengidentifikasi informasi berguna dari basis data berskala besar. Penerapan data mining yang tepat terhadap dataset medis berskala besar sangat penting untuk menghasilkan wawasan yang bermakna bagi pengambilan keputusan berbasis data.

Salah satu contoh dataset medis berskala besar adalah Imaging Data Commons (IDC) yang dikelola oleh National Cancer Institute (NCI). IDC adalah repositori data citra medis kanker berskala besar dan mencerminkan karakteristik utama Big Data secara nyata. Dataset ini menyimpan lebih dari 46,8 juta citra medis dalam format DICOM yang mencakup lebih dari 50 modalitas pencitraan, seperti CT scan, MRI, PET scan, dan X-ray, yang bersumber dari berbagai institusi penelitian kanker terkemuka di seluruh dunia dan dapat diakses melalui Google BigQuery. Efendi & Fatah (2025) menunjukkan bahwa pendekatan data mining berbasis clustering terbukti efektif memberikan wawasan mendalam tentang distribusi pada dataset berukuran besar.

Algoritma K-Means merupakan teknik clustering berbasis Unsupervised Learning yang paling banyak digunakan dalam data mining karena metode ini mampu mengelompokkan data berdasarkan tingkat kemiripan karakteristik tanpa memerlukan label data sebelumnya. Hidayat et al. (2023) membuktikan bahwa algoritma K-Means mampu melakukan pengelompokan data berjumlah besar dalam waktu komputasi yang cepat dan efisien, serta menghasilkan nilai silhouette coefficient tertinggi dibandingkan algoritma clustering lain seperti Self Organizing Maps, DBSCAN, dan Louvain Clustering. Bahkan di tingkat yang lebih canggih, Salloum (2025) mendemonstrasikan bahwa kombinasi K-Means clustering dengan CNN dan ekstraksi fitur HOG pada 2.788 citra kanker payudara menghasilkan akurasi klasifikasi sebesar 98%, membuktikan bahwa pengelompokan berbasis K-Means mampu mengungkap struktur inheren data citra medis sebelum analisis supervised diterapkan.

Meskipun demikian, algoritma ini memiliki keterbatasan berupa sensitivitas terhadap pencilan (outlier) dan keharusan menentukan nilai K secara manual sebelum proses clustering dimulai. Dalam penelitian ini, keterbatasan tersebut diantisipasi melalui normalisasi MinMaxScaler, penyaringan modalitas tidak representatif, serta penentuan K optimal secara objektif menggunakan Elbow Method.

Meskipun penerapan K-Means dalam bidang kesehatan telah banyak dilakukan, terdapat kesenjangan penelitian yang signifikan. Penelitian terdahulu umumnya berfokus pada segmentasi citra pada level piksel, pengelompokan data rekam medis pasien, serta pengelompokan wilayah geografis berdasarkan risiko kesehatan. Ketiga arah tersebut memiliki

kesamaan mendasar, yaitu objek analisisnya adalah konten citra atau data pasien, bukan metadata distribusi modalitas itu sendiri. Namun, belum ada penelitian yang secara spesifik menganalisis pola distribusi dan pengelompokan metadata modalitas pencitraan medis pada repositori skala petabyte seperti IDC menggunakan pendekatan data mining berbasis SQL BigQuery. Padahal, data nyata yang diperoleh langsung dari BigQuery (Tabel 1) memperlihatkan realita yang cukup mengkhawatirkan, hanya dua modalitas CT dan MR sudah menguasai lebih dari 94% dari seluruh citra yang terindeks, sementara puluhan modalitas lain harus berbagi sisa di bawah 6%. Ketiadaan pemetaan distribusi yang terstruktur ini berpotensi menyebabkan pengembang model AI medis mengabaikan risiko bias data sejak tahap awal pelatihan model.

Berdasarkan kesenjangan tersebut, penelitian ini mengajukan pertanyaan penelitian: Bagaimana pola pengelompokan modalitas pencitraan medis dalam dataset IDC dapat diidentifikasi secara sistematis menggunakan K-Means Clustering berbasis metadata BigQuery, dan apa implikasinya terhadap pengembangan model kecerdasan buatan di bidang onkologi? Berbeda dari penelitian sebelumnya yang menganalisis konten citra, penelitian ini secara khusus menargetkan metadata distribusi modalitas yang diekstraksi langsung melalui kueri SQL pada Google BigQuery, sehingga menghasilkan pemetaan distribusi yang mencerminkan kondisi repositori secara nyata dan dapat dijadikan landasan mitigasi bias sebelum proses pelatihan model dimulai.

Tabel 1. Data Distribusi Modalitas Pencitraan dalam Dataset IDC

No	Modalitas	Jumlah Gambar	Persentase (%)	Body Part Unik
1	CT	29.163.883	62,22%	69
2	MR	15.189.444	32,41%	39
3	PT	1.338.339	2,86%	15
4	SM	365.048	0,78%	0
5	SR	270.687	0,58%	0
6	MG	268.365	0,57%	3
7	SEG	214.182	0,46%	33
8	Lainnya (40+ jenis)	~282.000	< 0,12% *each*	beragam
	TOTAL	≥ 47.091.948	≈ 100%	69+ jenis

Data pada Tabel 1 menunjukkan dari total citra yang terindeks, dua modalitas saja, CT dan MR sudah menguasai lebih dari 94,6% keseluruhan data. CT bahkan mencakup 69 jenis body part unik dan MR mencakup 39 jenis, menunjukkan cakupan klinis yang luar biasa luas. Sebaliknya, SM dan SR masing-masing tercatat dengan unique_body_parts = 0, artinya tidak memiliki data anatomi tubuh yang terdefinisi. Ketimpangan distribusi ini menjadi permasalahan utama yang perlu dipetakan melalui pendekatan clustering K-Means agar diperoleh pemahaman lebih terstruktur tentang bagaimana data modalitas tersebar, dan apa implikasinya bagi pengembangan riset kanker berbasis pencitraan medis ke depan.

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk: (1) mengidentifikasi kelompok (cluster) modalitas pencitraan medis berdasarkan karakteristik distribusi data dalam dataset IDC; (2) menerapkan tahapan preprocessing yang tepat pada metadata DICOM yang diakses melalui Google BigQuery; (3) mengimplementasikan algoritma K-Means menggunakan Python dengan penentuan K optimal melalui Elbow Method; serta (4) mengevaluasi kualitas cluster menggunakan Silhouette Score dan mengekstraksi wawasan bermakna dari hasil clustering untuk mendukung pengembangan penelitian onkologi berbasis data.

2. METODOLOGI

a) Sumber Data

Data penelitian diperoleh dari Google BigQuery Public Dataset dengan spesifikasi sebagai berikut:

Tabel 2. Spesifikasi Sumber Data

Parameter	Keterangan
Platform	Google BigQuery
Project	bigquery-public-data
Dataset	idc_current
Tabel	dicom_metadata
Jumlah data awal	26 modalitas
Format output	CSV (idc_variety.csv)

Pengambilan data dilakukan menggunakan kueri SQL berikut pada Google BigQuery:

```
SELECT
Modality,
COUNT(*) AS image_count,
ROUND(COUNT(*) * 100.0 / SUM(COUNT(*))
OVER(), 2) AS percentage,
COUNT(DISTINCT BodyPartExamined) AS
unique_body_parts
FROM `bigquery-public-
data.idc_current.dicom_metadata`
GROUP BY Modality
ORDER BY image_count DESC
```

b) Variabel Penelitian

Variabel yang digunakan dalam proses *clustering* adalah sebagai berikut:

Tabel 3. Variabel/Fitur yang Digunakan

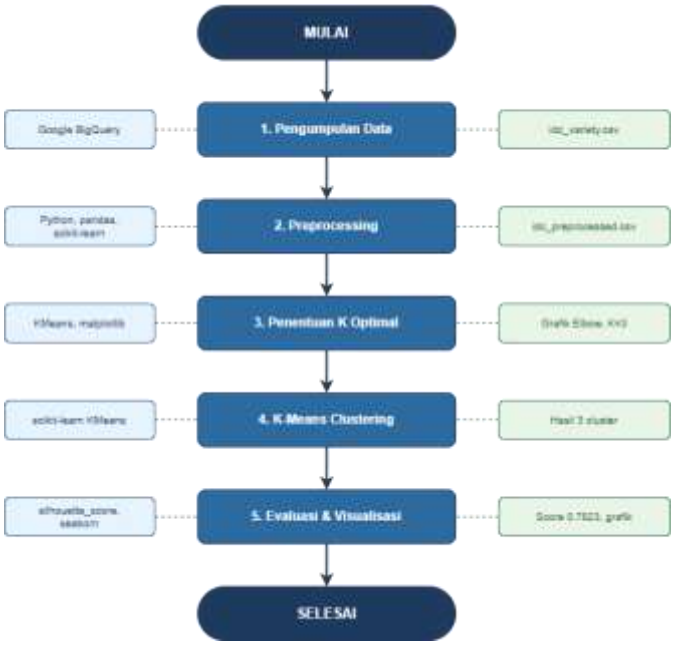
Kolom	Deskripsi	Tipe	Peran
Modality	Jenis modalitas imaging	String	Label identitas
image_count	Jumlah total gambar	Integer	Fitur utama (F1)

percentage	Persentase terhadap total	Float	Fitur utama (F2)
unique_body_parts	Jumlah body part unik	Integer	Fitur utama (F3)

Ketiga fitur tersebut dipilih karena secara bersama-sama merepresentasikan tiga dimensi karakteristik modalitas yang saling melengkapi: image_count mencerminkan volume penggunaan aktual, percentage merepresentasikan proporsi relatif terhadap keseluruhan dataset, dan unique_body_parts menggambarkan keragaman cakupan klinis dari setiap modalitas. Kombinasi ketiganya memungkinkan algoritma K-Means membedakan modalitas tidak hanya berdasarkan seberapa banyak datanya, tetapi juga seberapa luas relevansi klinisnya.

c) Alur Penelitian

Penelitian ini mengikuti alur KDD (*Knowledge Discovery in Databases*) dengan tahapan sebagai berikut:



Gambar 1. Alur Metodologi Penelitian

Secara naratif, penelitian diawali dengan pengumpulan data melalui kueri SQL pada Google BigQuery untuk mengekstrak metadata distribusi modalitas dari tabel dicom_metadata. Data hasil kueri kemudian memasuki tahap preprocessing yang mencakup pembersihan nilai null, penghapusan duplikat, penyaringan noise, dan normalisasi fitur. Selanjutnya, data yang telah bersih diproses menggunakan algoritma K-Means Clustering dengan penentuan nilai K optimal melalui Elbow Method. Tahap akhir adalah evaluasi kualitas clustering menggunakan Silhouette Score dan interpretasi hasil untuk menghasilkan wawasan yang bermakna secara ilmiah.

d) Tahap Preprocessing

Preprocessing dilakukan menggunakan Python dengan *library* pandas dan scikit-learn, meliputi tahapan berikut:

1. Penghapusan nilai *null*: menghapus baris dengan nilai kosong pada kolom Modality, image_count, atau unique_body_parts.
2. Penghapusan duplikat: menghapus baris yang memiliki nilai Modality yang sama.
3. *Filtering noise*: menghapus modalitas dengan image_count < 100 karena modalitas dengan jumlah citra di bawah ambang batas tersebut dianggap tidak representatif secara statistik dan berpotensi menjadi pencilan (outlier) yang mengganggu kestabilan centroid K-Means.
4. Normalisasi MinMaxScaler: dipilih karena rentang nilai antar fitur sangat timpang — image_count bernilai hingga puluhan juta sementara unique_body_parts hanya puluhan. Tanpa normalisasi, fitur dengan skala besar akan mendominasi perhitungan jarak Euclidean sehingga fitur lain kehilangan pengaruhnya. MinMaxScaler mentransformasi seluruh nilai ke rentang [0, 1] secara proporsional sehingga setiap fitur berkontribusi secara seimbang dalam proses clustering

3. HASIL DAN PEMBAHASAN

a) Konfirmasi Unsupervised Learning

Sebelum analisis dimulai, dilakukan konfirmasi terhadap jenis pendekatan *learning* yang sesuai. Dari inspeksi struktur tabel dicom_metadata di BigQuery, dikonfirmasi bahwa tidak ditemukan satu pun kolom yang berfungsi sebagai label prediksi. Seluruh kolom bersifat deskriptif. Dengan demikian, pendekatan *Unsupervised Learning*

dengan *K-Means Clustering* adalah satu-satunya pendekatan yang valid dan tepat untuk dataset ini.

b) Hasil Pengumpulan Data

Query SQL dijalankan pada tabel bigquery-public-data.idc_current.dicom_metadata untuk mengambil data distribusi modalitas. Berikut adalah ringkasan data yang diperoleh:

Tabel 4. Data Distribusi Modalitas IDC (Top 10)

Modalitas	Jumlah Gambar	Persentase (%)	Body Parts Unik
CT	29.163.883	62,22	69
MR	15.189.444	32,41	39
PT	1.338.339	2,86	15
SM	365.048	0,78	0
SR	270.687	0,58	0
MG	268.365	0,57	3
SEG	214.182	0,46	33
XC	14.124	0,03	1
CR	12.427	0,03	19
ANN	7.108	0,02	0

Total keseluruhan data mencakup 26 modalitas sebelum *preprocessing*. Terlihat jelas dominasi CT (62,22%) dan MR (32,41%) yang bersama-sama mencakup 94,63% dari total dataset.

c) *Hasil Preprocessing*

Setelah melalui tahap *preprocessing*, data mengalami beberapa perubahan sebagai berikut:

Tabel 5. Perbandingan Data Sebelum dan Sesudah Preprocessing

Proses	Jumlah Baris	Keterangan
Data awal	26	Semua modalitas dari BigQuery
Setelah hapus null	26	Tidak ada nilai null
Setelah hapus duplikat	26	Tidak ada duplikat Modality
Setelah filter noise (<100)	21	5 modalitas dengan data sangat kecil dibuang
Setelah normalisasi	21	Nilai dalam rentang [0.0, 1.0]

Normalisasi MinMaxScaler menghasilkan nilai CT = (1.0, 1.0, 1.0) sebagai nilai tertinggi yang menjadi patokan skala, sedangkan modalitas lainnya memiliki nilai yang sangat kecil mencerminkan disparitas distribusi data yang ekstrem.

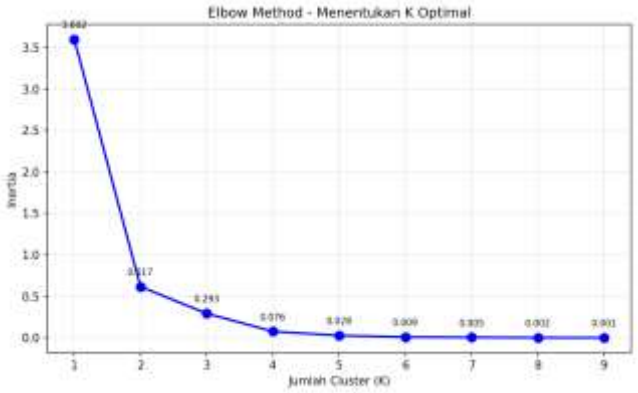
d) Hasil Elbow Method

Elbow Method dijalankan untuk nilai K dari 1 hingga 9. Berikut adalah hasil perhitungan *inertia*:

Tabel 6. Hasil Elbow Method (Inertia per Nilai K)

K (Jumlah Cluster)	Inertia	Keterangan
1	3.6017	Semua data dalam satu cluster
2	0.6170	Penurunan sangat besar
3	0.2929	Penurunan masih signifikan (TITIK SIKU)
4	0.0765	Mulai landai
5	0.0280	Penurunan kecil
6	0.0092	Hampir datar
7	0.0055	Sangat kecil
8	0.0020	Tidak signifikan
9	0.0011	Minimal

Berdasarkan grafik *Elbow Method*, titik siku yang paling jelas terlihat pada K=3, di mana penurunan inertia dari K=2 ke K=3 masih cukup signifikan (dari 0,617 menjadi 0,293), namun mulai melambat secara drastis setelah K=3. Oleh karena itu, K=3 ditetapkan sebagai jumlah *cluster* optimal.



Gambar 2. Grafik Elbow Method

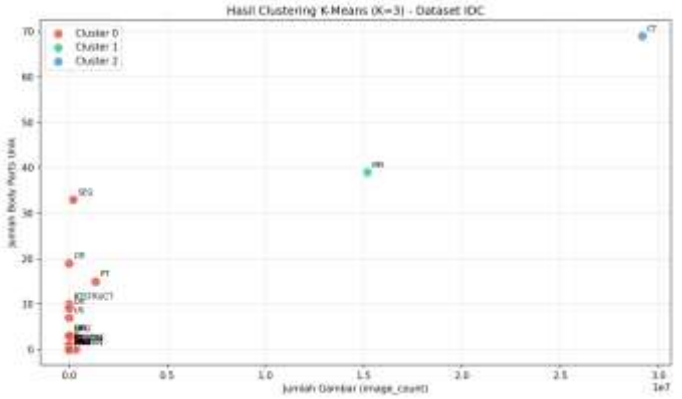
e) Hasil K-Means Clustering

Dengan K=3, algoritma *K-Means* menghasilkan pengelompokan sebagai berikut:

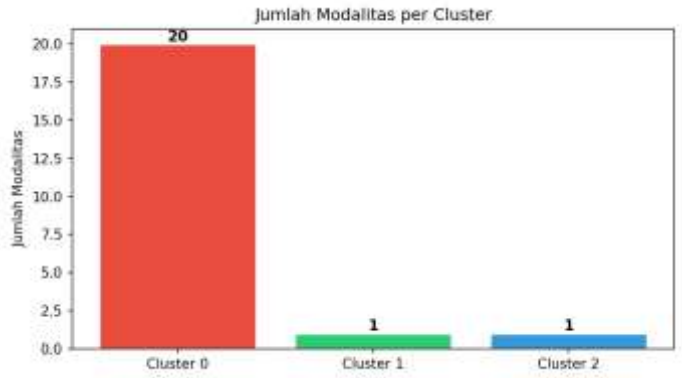
Tabel 7. Hasil Clustering K-Means (K=3)

Cluster	Anggota	Total Gambar	%
Cluster 2	CT	29.163.883	62,22%
Cluster 1	MR	15.189.444	32,41%
Cluster 0	PT, SM, SR, MG, SEG, XC, CR, ANN, US, RTSTRUCT, FUSION,	2.296.576	5,37%

DX, M3D, OT, PR, SC, RTDOSE, RTPLAN, XA, NM (20 modalitas)



Gambar 3. Scatter Plot Hasil Clustering



Gambar 4. Bar Chart Jumlah Modalitas per Cluster

f) Evaluasi Kualitas Clustering

Kualitas hasil *clustering* diukur menggunakan *Silhouette Score* yang menghasilkan nilai 0,7823. Nilai ini termasuk dalam kategori *clustering* berkualitas tinggi (nilai > 0,7), menunjukkan bahwa pembagian data ke dalam 3 *cluster* sudah sangat sesuai dan setiap modalitas berada dalam *cluster* yang paling tepat untuk karakteristiknya.

Tabel 8. Interpretasi Nilai Silhouette Score

Rentang Nilai	Interpretasi	Hasil Penelitian Ini
0.71 – 1.00	Struktur clustering kuat/sangat baik	✓ 0.7823 Sangat Baik
0.51 – 0.70	Struktur clustering sedang/cukup baik	—
0.26 – 0.50	Struktur clustering lemah	—
< 0.25	Tidak ada struktur clustering	—

g) Pembahasan

Hasil *clustering* K-Means mengungkap sejumlah informasi penting yang tidak dapat ditemukan hanya dengan membaca tabel distribusi biasa. Temuan ini sejalan dengan Efendi & Fatah (2025) yang membuktikan bahwa pendekatan *data mining* berbasis *clustering* efektif memberikan wawasan mendalam tentang distribusi pada dataset berukuran besar yang tidak dapat diperoleh melalui analisis konvensional.

Pertama, ditemukan hubungan eksponensial antara seberapa banyak bagian tubuh yang bisa dicitrakan oleh suatu modalitas dengan seberapa sering modalitas itu digunakan. CT yang mencakup 69 *body parts* memiliki 29 juta gambar, MRI dengan 39 *body parts* memiliki 15 juta gambar. Namun PT yang mencakup 15 *body parts* hanya memiliki 1,3 juta gambar, bukan sepertiga dari CT, melainkan hanya 4,6%-nya. Semakin serbaguna sebuah modalitas, pertumbuhan penggunaannya jauh melampaui proporsi keserbagunaan itu sendiri. Implikasinya, pengembang AI medis sebaiknya memprioritaskan CT dan MRI bukan hanya karena datanya banyak, tetapi karena model yang dilatih pada dua modalitas ini secara otomatis mencakup spektrum diagnosis yang jauh lebih luas. Satu model CT yang baik berpotensi berguna untuk 69 jenis kondisi klinis sekaligus.

Kedua, K-Means menempatkan PT, MG, dan RTPLAN dalam satu *cluster* yang sama. Bukan karena fungsi klinisnya mirip, tetapi karena karakteristik datanya identik *volume* kecil dan cakupan *body part* terbatas. Dari perspektif klinis ketiganya berbeda jauh, namun dari perspektif *data science* ketiganya menghadapi tantangan yang sama yaitu kekurangan data. Hal ini membuktikan kemampuan K-Means sebagai teknik *unsupervised learning* yang, sebagaimana ditegaskan oleh Ramadhani et al. (2025), bekerja tanpa memerlukan label data namun tetap mampu menghasilkan pengelompokan bermakna pada data medis. Pola serupa juga ditemukan oleh Hidayat et al. (2023) dalam *clustering* daerah rawan stunting di Jawa Barat, di mana wilayah yang secara geografis berbeda dikelompokkan bersama karena kesamaan

karakteristik datanya. Implikasinya, solusi untuk meningkatkan kualitas AI pada satu modalitas minor misalnya teknik *data augmentation* untuk MG kemungkinan besar dapat langsung diterapkan untuk semua 20 modalitas lainnya tanpa perlu pendekatan yang berbeda-beda.

Hasil evaluasi menggunakan *Silhouette Score* menunjukkan nilai 0,7823 yang masuk dalam kategori *cluster* kuat (0,71–1,00). Nilai ini lebih tinggi dibandingkan penelitian Hidayat et al. (2023) yang memperoleh nilai 0,61 dengan kategori *cluster* layak menggunakan dataset yang berbeda. Perbedaan ini menunjukkan bahwa karakteristik distribusi modalitas IDC memiliki pemisahan antar kelompok yang lebih tegas, mencerminkan ketimpangan distribusi data yang sangat ekstrem antara CT, MRI, dan modalitas lainnya.

Jika dibandingkan dengan penelitian sejenis, hasil penelitian ini menunjukkan konsistensi sekaligus kekhasan tersendiri. Efendi & Fatah (2025) yang menerapkan K-Means pada data penyebaran Covid-19 juga berhasil mengidentifikasi kelompok wilayah berdasarkan intensitas kasus, namun objek analisisnya adalah data epidemiologis bukan metadata repositori citra medis. Mahmudah et al. (2023) menggunakan K-Means untuk mengelompokkan wilayah rawan stunting dan menghasilkan tiga cluster bermakna, serupa dengan hasil K=3 dalam penelitian ini, meskipun konteks datanya berbeda. Sementara itu, Jiang (2024) menunjukkan bahwa K-Means tetap efisien pada dataset berskala besar, yang relevan mengingat data IDC mencakup puluhan juta citra. Perbedaan mendasar penelitian ini dibandingkan seluruh referensi tersebut terletak pada objek clustering-nya, penelitian ini menganalisis distribusi metadata modalitas pada repositori petabyte, suatu pendekatan yang belum pernah dilakukan sebelumnya pada dataset IDC.

4. KESIMPULAN

Penelitian ini berhasil menerapkan algoritma *K-Means Clustering* pada *dataset Imaging Data Commons* (IDC) yang diakses melalui Google BigQuery untuk mengelompokkan modalitas pencitraan medis kanker. Melalui tahapan *preprocessing* yang mencakup pembersihan data, penyaringan *noise*, dan normalisasi MinMaxScaler, serta penentuan K optimal menggunakan *Elbow Method*, terbentuk tiga *cluster* yang bermakna secara ilmiah dengan *Silhouette Score* 0,7823 yang termasuk kategori *cluster* kuat. *Cluster* 2 ditempati CT (62,22%), *Cluster* 1 oleh MR (32,41%), dan *Cluster* 0 menghimpun 20 modalitas minor dengan total 5,37% data.

Lebih dari sekadar pengelompokan teknis, penelitian ini mengungkap temuan yang tidak dapat diperoleh dari pembacaan tabel biasa. Hubungan antara cakupan *body part* dan *volume* penggunaan modalitas bersifat eksponensial, dua puluh modalitas yang berbeda secara klinis ternyata menghadapi tantangan *data science* yang identik sehingga satu strategi penanganan dapat diterapkan sekaligus, serta dominasi CT dan MR sebesar 94,63% berpotensi menciptakan *blind spot* pada model kecerdasan buatan medis apabila ketimpangan ini tidak ditangani sebelum proses pelatihan. Temuan-temuan ini menegaskan nilai strategis pendekatan *data mining* berbasis *clustering* dalam mengekstrak pengetahuan bermakna dari repositori pencitraan medis berskala *petabyte* guna mendukung pengembangan riset onkologi berbasis data yang lebih terarah.

DAFTAR PUSTAKA

Aditya, B., Harjanta, A. T. J., & Latifah, K. (2025). Implementasi algoritma k-means clustering pada sistem informasi geografis fasilitas kesehatan BPJS Kesehatan Kota Semarang. *Jurnal Informatika Teknologi dan Sains (JINTEKS)*, *7*(1), 438-448.

Alzakari, S. A., Alruwais, N., Sorour, S., Ebad, S. A., Elnour, A. A. H., & Sayed, A. (2024). A big data analysis algorithm for massive sensor medical images. *PeerJ Computer Science*, *10*, e2464.

Aulia, W., Siahaan, A. P. U., Marlina, L., Khairul, & Iqbal, M. (2024). K-means clustering algorithm analysis for grouping patient medical record data based on disease type.

Jurnal Info Sains: Informatika dan Sains, *14*(04).

Awad, F. H., Hamad, M. M., & Alzubaidi, L. (2023). Robust classification and detection of big medical data using advanced parallel k-means clustering, YOLOv4, and logistic regression. *Life*, *13*(3), 691

Bili, M. L., Mau, S. D. I., & Momo, L. L. (2025). Opini Masyarakat terhadap infrastruktur Desa Malitidari menggunakan metode *unsupervised learning* pada tools Orange. *Modem: Jurnal Informatika dan Sains Teknologi*, *3*(4), 57–67.

Efendi, M. A., & Fatah, Z. (2025). Penerapan data mining untuk mengelompokkan penyebaran Covid-19 di Indonesia menggunakan algoritma k-means. *JAMASTIKA*, *4*(1).

Fedorov, A., & Homeyer, A. (2023). The NCI Imaging Data Commons as a platform for reproducible research in computational pathology. *Computer Methods and Programs in Biomedicine, 242*, 107839.

González García, C., & Álvarez-Fernández, E. (2022). What is (not) Big Data based on its 7Vs challenges: A survey. *Big Data and Cognitive Computing*, *6*(4), 158

Hidayat, T., Jajuli, M., & Susilawati. (2023). Clustering daerah rawan stunting di Jawa Barat menggunakan algoritma k-means. *INFOTECH: Jurnal Informatika & Teknologi*, *4*(2), 137-146.

IBM. (2021). What is data mining? IBM. https://www.ibm.com/topics/data-mining

Indraputra, R. A., & Fitriana, R. (2020). K-Means clustering data COVID-19. *Jurnal Teknik Industri*, *10*(3), 275–282.

Jiang, Z. (2024). Research on performance optimization of k-means algorithm on large dataset. *International Journal of Advance in Applied Science Research*, *3*.

Mahmudah, F., Rahaningsih, N., Dana, R. D., & Rohmat, C. L. (2023). Implementasi data mining menggunakan algoritma k-means untuk mempermudah pengelompokkan wilayah rawan stunting di Kabupaten Cirebon. *Jurnal INTEK (Informatika dan Teknologi Informasi)*, *8*(1), 44-52.

McCarthy, N., Dahlan, A., Cook, T. S., O'Hare, N., Ryan, M. L., St John, B., ... & Curran, K. M. (2021). Enterprise imaging and big data: A review from a medical physics perspective. *Physica Medica*, *83*, 206-220.

Ramadhani, M. R., Hermawan, R. N., Fajrian, I., Rachmat, D. A., Sumanto, & Kuswanto, A. D. (2025). Analisis klaster pasien diabetes menggunakan algoritma k-means berdasarkan usia, kadar glukosa, dan tekanan darah. *RJTI (Riau Jurnal Teknik Informatika)*, *4*(2), 374-378.

Salloum, S. (2025). K-means clustering and classification of breast cancer images using histogram of oriented gradients features and convolutional neural network models: Diagnostic image analysis study. *JMIR Formative Research*, *9*, e71974.

Schacherer, D. P., Herrmann, M. D., Clunie, D. A., Höfener, H., Clifford, W., Longabaugh, W. J. R., Pieper, S., Kikinis, R., Fedorov, A., & Homeyer, A. (2023). The NCI Imaging Data Commons as a platform for reproducible research in computational pathology. Computer Methods and Programs in Biomedicine, 242, 107839.

Suwanto, F., & Pradesan, I. (2026). Implementasi data mining regresi linear berganda pada sistem prediksi penjualan obat pada apotek XYZ. *Jurnal Sistem Informasi TGD, 5*(1), 83–94.

Utomo, Y. B., Kurniasari, I., & Yanuartanti, I. (2023). Penerapan knowledge discovery in database untuk analisa tingkat kecelakaan lalu lintas. *Jurnal Teknik Informatika Kaputama (JTIK)*, *7*(1), 171–1